An Introduction to AI for GLAM

Anonymous Authors[1]

Abstract

There is a growing interest in the utilisation of Machine Learning (ML) techniques within Galleries, Libraries, Archives and Museums (GLAM), and a corresponding demand for training, to enable practitioners to engage confidently in this area. Staff at these institutions are seeking practical knowledge and skills in machine learning concepts and methods, specific to the work of the sector, such as in the curation and collection of heritage collections. In this paper we discuss the motivations and methods behind 'An Introduction to AI for GLAM', a new Carpentries (Car, b). workshop under development through an international partnership between British Library, Smithsonian Institution, and The National Archives UK. This new workshop aims to introduce GLAM practitioners to the key conceptual and practical considerations for supporting, participating in and undertaking machine learning-based research and projects within the GLAM sector.

1. Introduction

The past decade has seen a growing exploration of machine learning by the Galleries, Libraries, Archives and Museums (GLAM) sector. This interest is reflected in a growing number of networking initiatives[1], research projects applying machine learning to GLAM collections (Lee et al., 2020; Lincoln et al., 2020; Dee) and projects developing machine learning tools aimed explicitly at GLAM institutions (Kahle et al., 2017).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Teaching Machine Learning Workshop at ECML 2021. Do not distribute.

[1]These include AI4LAM: https://sites.google.com/view/ai4lam, CENL AI in Libraries Network Group: https://www.cenl.org/networkgroups/ai-in-libraries-network-group/, and the AEOLIAN network https://www.aeolian-network.net/about/

A potential barrier to the effective adoption of machine learning within the sector is a "skills gap" amongst staff (Cox, 2021). The exact content of this skills gap depends largely on the perceived role of GLAM staff in relation to Machine Learning. Should staff be directly building and training models? Should they be able to document training data developed from library content? Should they be able to work as part of a team developing machine learning models?

This paper briefly provides some background on the GLAM sector and existing training initiatives aimed at GLAM staff. It then introduces our Carpentries workshop 'An Introduction to AI for GLAM'.

2. Machine Learning and the GLAM Sector

The GLAM sector encompasses a broad range of institutions in terms of size, budget, collections and primary audience. It is therefore difficult to make broad statements about the sector that will be completely accurate. However, there are areas of common activity and focus:

- Cataloguing and other forms of metadata generation.
- Enabling search and discovery of collections.
- Supporting and carrying out research.
- Public engagement and crowdsourcing.

These areas are all ones in which machine learning could be – or already is – having an impact. Beyond this there are also shared aims across the GLAM sector as well as existing groups organised as part of the GLAM sector (Zorich et al., 2008). This suggests that developing teaching material for this sector is worthwhile.

3. GLAM driven ML Training Initiatives

Major GLAM institutions such as the British Library, The National Archives UK and the Smithsonian Institution have all undertaken to develop a wide range of training opportunities for their staff in this area, taking a variety of training approaches in getting this complex material across to a diversity of staff.

3.1. Library Carpentry

Library Carpentry is an initiative under the Carpentries (Car, b) umbrella focused on teaching software skills to library and information related communities. Library Carpentry was originally developed in 2014 in response to demand for access to sofware and data skills amongst people working in libraries (Baker et al., 2016). The Library Carpentry curriculum covers materials that would be relevant to GLAM staff wanting to work with machine learning, including introductions to 'tidy data', version control systems, use of the command line, and basic programming in Python. [2]

3.2. Digital Scholarship Training Programme at British Library

The Digital Research and Curator Team at the National Library of the UK, has run a digital and data skills training programme for British Library staff since 2012 (Dig). This programme creates opportunities for colleagues to develop necessary skills and knowledge to support, and undertake in their own right, emerging areas of modern scholarship such as the Digital Humanities. The programme supports several different modes of training delivering, from one-hour lunchtime lectures and a monthly reading group, 2 hour hands-on exploratory Hack and Yacks, to full day or week long courses on a given digital topic. The topic of machine learning has featured heavily throughout this programme in recent years, with a particular emphasis on introducing high level concepts and use cases in the application of ML in the GLAM sector to a novice audience, rather than delivering the software skills to implement them.

3.3. Machine Learning Club at The National Archives

The National Archives have run a number of activities under the banner of Machine Learning Club for all interested staff. The initiative began with a one-off introductory talk, expanding to a series of talks covering multiple aspects of the topic. Enthusiasm amongst attendees to learn more led to the running of a series of workshops for technical and non-technical staff alike to gain hands-on experience of machine learning. As with the British Library, the aim was to teach concepts and for participants to understand their role within an ML eco-system.

3.4. Computing for Cultural Heritage

The Computing for Cultural Heritage project (Com, a) was an Institute of Coding funded trial (2019/2021) that saw Birkbeck University of London, British Library and The National Archives UK develop a new one-year postgraduate certificate aimed at providing information professionals across GLAM sector with an understanding of basic programming, analytic tools and computing environments to support them in their daily work. It was born out of a need to further support those who, having gained a keen interest in machine learning through institutional staff training, required the programming skills to practically undertake it(Com, b).

These diverse training initiatives, all designed from a GLAM practitioners perspective on ML, have had positive impacts, with participants reporting increased confidence in interactions with external data scientists, and being able to bring ML into their own research.[3]

4. An Introduction to AI for GLAM

In this section, we introduce our workshop, "An Introduction to AI for GLAM", currently under development as part of the "Carpentries Incubator".[4] This next section outlines the aims, topics and delivery methods of this workshop alongside a reflection on why these were chosen.

The material for this course were developed as part of a Carpentries Lesson Development Study Group (Car, a). This study group took place over a couple of months and was intended to help participants develop new Carpentries lessons. Four members of the group, representing three GLAM institutions, decided to collaborate on this AI lesson.

4.1. Learner profiles

As part of the process of developing the lesson materials, learner profiles were created. The Carpentries recommend the development of learner profiles as a way of better identifying the target audience and their needs (Wilson, 2019). Learner profiles require the description of some characteristics of an expected learner for the material. For example; "What is their expected educational level?", "What type of exposure do they have to the technologies you plan to teach?" and "What are the pain points they are currently experiencing?" (Wilson, 2019). Although the people depicted in the profiles are fictional, the development of the profiles often drew on real people who work in our home institutions.

One challenge of using the learner profiles for our lesson was adapting them to focus slightly less on their technical skills. The learner profiles were also adapted slightly to focus more on our learners' potential attitudes toward machine learning. Whilst we were developing the lessons, we

[2] https://librarycarpentry.org/lessons/

[3] See for instance student projects undertaken by Computing for Cultural Heritage https://www.bl.uk/case-studies/computing-for-cultural-heritage-student-projects and https://blog.nationalarchives.gov.uk/computing-cholera-topic-modelling-the-catalogue-entries-

[4] The incubator is a place for lessons to be developed outside of the core Carpentries curriculum https://github.com/carpentries-incubator

wanted to keep in mind the varying levels of enthusiasm for machine learning from the target audience.

4.2. Aims

The role that GLAM staff should play in machine learning projects remains an open question. Some GLAM institutions might choose to "outsource" most of their machine learning efforts to commercial solutions, whilst others will want their staff to be involved more directly. Whilst the stakes are not as high, this mirrors discussions in other disciplines, such as medicine, around the role domain experts should play and, crucially, what they need to know about machine learning (Sim et al., 2021; Olczak et al., 2021). We would argue that regardless of whether GLAM staff will be "directly" involved, machine learning methods will be underpinning so many technologies in the future that basic literacy around machine learning is crucial for all GLAM staff.

Following this desire to develop a basic understanding of machine learning, 'An Introduction to AI for GLAM' has several aims. The primary overarching goal of the material is to provide an accessible introduction to machine learning for GLAM staff that is relevant to their work. In particular, the materials aim to introduce basic machine learning concepts and demystify training machine-learning models. The material also seeks to provide a high-level overview of what machine learning is good at and where its limitations lie. Giving a realistic account of the field is especially important in the context of the growing use of machine learning in the GLAM sector and the risks of machine learning approaches being "oversold" to staff.

Another central aim of the course is to emphasise ethical considerations. There are growing calls and examples of integrating ethics into the curriculums of machine learning courses (Garrett et al., 2020; Saltz et al., 2019). Highlighting potential ethical issues, particularly related to GLAM "data" challenges, is a possible role for GLAM staff in machine learning (Coleman, 2020).

Finally, the lesson material aims to give a sense of the steps involved in a machine learning project, from identifying a "business need" to deploying and monitoring models. The overview of these steps aims to help prepare GLAM staff to work as part of a machine learning project team rather than on the technical implementation of these different stages of a machine learning project.

4.3. Lesson Topics

We chose the topics covered in the workshop material with the learners' and aims in mind. The topics, as a result, differ from what may often be covered as part of an introductory machine learning course. The current "episodes" include:

- What are Artificial Intelligence (AI) and Machine Learning?
- What is Machine Learning good at?
- Understanding and managing bias.
- Applying Machine Learning.
- The Machine Learning ecosystem.

Since the aims of the session are to provide a broad introduction to the topic for GLAM staff, the topics have a strong focus on the practical applications and steps in machine learning and a relatively more minor focus on some conceptual topics. Some topics included in most machine learning introductions, for example, loss functions, are not covered in any detail. With that said, the first complete module of the session does provide a basic conceptual introduction to machine learning.

4.4. Delivery Methods

Unlike many introductions to machine learning, the materials do not include 'hands-on' programming components. However, the goal of the lesson material is still to be practically focused and interactive. The Carpentries place a strong emphasis on interactivity and the inclusions of exercises in lesson materials.

Since the lessons do not cover coding, we did not include coding exercises. Instead, the aims of the exercises were to:

- Test and develop an understanding of important machine learning concepts.
- Encourage a reflection of how machine learning could be used within a GLAM setting.
- Encourage a reflection on how machine learning could be utilised within specific institutions in which the participants of the workshops are based.
- Emphasise a critical reflection on the ethical issues that can be raised by using machine learning and particularly how these appear in the context of GLAM institutions and collections.

Examples of the types of exercises include;

- Multiple choice questions designed to test that concepts have been understood. For example, understanding the difference between supervised and unsupervised learning.
- A Discussion prompt for thinking about how machine learning could be utilised: "How might object detection help speed up digitisation?".

- Group discussion of "points at which bias may enter the pipeline, and questions/strategies GLAM staff might want to consider in order to manage it."
- A "hands-on" activity exploring commercial computer vision services to reflect on their potential strengths and weaknesses.

We aimed for these exercises to build on experiences of running Library Carpentry workshops which includes similar discussion exercises.[5] These types of exercise are often not focused on teaching a "hands-on" skill as such but instead focus on increasing the learners broad understanding of a topic and, crucially, their confidence.

4.5. Community development and maintenance

We have several reasons for developing this workshop as part of the Carpentries ecosystem. The Carpentries workshop materials are generated from source files in a GitHub repository. This allows for anyone to make a pull request or open an issue related to the lesson. Beyond this technical ability to make changes, the Carpentries organise regular events to encourage review and development of existing materials.

Machine learning is a rapidly developing field with regular technical advances. Beyond this, there is also a growing maturity around the deployment of machine learning models in various domains. The use of machine learning in the GLAM sector will continue to develop over time, making it likely that aspects of our lesson will need to be updated. We hope that integrating the lesson materials inside the Carpentries ecosystem will help ensure that a broader community can update the material.

5. Conclusion

Introductory machine learning training that is grounded in the specific applications and use cases relevant to cultural heritage, that is practical, without being too overtly technical, will be key to ensuring the wider adoption of machine learning methods across GLAM. Though major cultural heritage institutions have undertaken in recent years to provide their own staff with a variety of training in this area, there is much to be gained by pooling expertise, resources, and experience to deliver a variety of open training materials available sector-wide. 'An Introduction to AI for GLAM' represents one effort, aiming to meet the growing demand for machine learning training specific to the sector, and provide a strong foundation for staff to gain confidence in entering this complex area. The Carpentries, and specifically Library Carpentry, though tending towards more technical lessons historically, provides the ideal home for such training. Library Carpentry acts as a natural sign-post for digital skill seeking GLAM professionals, and has a diverse and dedicated community at the ready to ensure not just the open, shared and continued maintenance of the materials, but the development of related and more advanced courses branching off the core foundation as necessary.

[5] An example includes a discussion exercise focused on "jargon-busting" terminology used around code or software development.

References

Carpentries Lesson Development Study Group. https://carpentries-incubator.github.io/study-groups/, a.

The Carpentries. https://carpentries.org/index.html, b.

Computing Cultural Heritage - The British Library. https://www.bl.uk/projects/computingculturalheritage, a.

Computing Cultural Heritage - student projects. https://www.bl.uk/case-studies/computing-for-cultural-heritage-student-projects, b.

Deep Discoveries. https://tanc-ahrc.github.io/DeepDiscoveries/index.html.

Digital Scholarship Training Programme. https://www.bl.uk/projects/digital-scholarship-training-programme.

Baker, J., Moore, C., Priego, E., Alegre, R., Cope, J., Price, L., Stephens, O., van Strien, D., and Wilson, G. Library Carpentry: Software skills training for library professionals. *LIBER Quarterly*, 26(3):141–162, November 2016. ISSN 2213-056X. doi: 10.18352/lq.10176.

Coleman, C. N. Managing bias when library collections become data. 2020.

Cox, A. The impact of ai, machine learning, automation and robotics on the information professions. Technical report, CILIP, 2021.

Garrett, N., Beard, N., and Fiesler, C. More Than "If Time Allows": The Role of Ethics in AI Education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 272–278, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375868.

Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pp. 19–24, November 2017. doi: 10.1109/ICDAR.2017.307.

Lee, B., Mears, J., Jakeway, E., Ferriter, M. M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., and Weld,

D. S. The newspaper navigator dataset: Extracting and analyzing visual content from 16 million historic newspaper pages in chronicling america. *ArXiv*, abs/2005.01583, 2020.

Lincoln, M., Corrin, J., Davis, E., and Weingart, S. B. Campi: Computer-aided metadata generation for photo archives initiative. 2020.

Olczak, J., Pavlopoulos, J., Prijs, J., Ijpma, F., Doornberg, J., Lundström, C., Hedlund, J., and Gordon, M. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a clinical ai research (cair) checklist proposal. *Acta orthopaedica*, pp. 1–13, 2021.

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N., and Beard, N. Integrating Ethics within Machine Learning Courses. *ACM Transactions on Computing Education*, 19(4):32:1–32:26, August 2019. doi: 10.1145/3341164.

Sim, J. Z. T., Fong, Q. W., Huang, W., and Tan, C. Machine learning in medicine: what clinicians should know. *Singapore medical journal*, 2021.

Wilson, G. *Teaching Tech Together: How to Make Your Lessons Work and Build a Teaching Community around Them*. CRC Press, October 2019. ISBN 978-1-00-072801-9.

Zorich, D., Waibel, G., and Erway, R. Beyond the Silos of the LAMs. *Collaboration Among Libraries*, 2008.